

# A VERSION OF SZEMERÉDI'S REGULARITY LEMMA FOR MULTICOLORED GRAPHS AND DIRECTED GRAPHS THAT IS SUITABLE FOR INDUCED GRAPHS

MARIA AXENOVICH AND RYAN MARTIN

**ABSTRACT.** In this manuscript we develop a version of Szemerédi's regularity lemma that is suitable for analyzing multicolorings of complete graphs and directed graphs. In this, we follow the proof of Alon, Fischer, Krivelevich and M. Szegedy [*Combinatorica* **20**(4) (2000) 451–476] who prove a similar result for graphs.

The purpose is to extend classical results on dense hereditary properties, such as the speed of the property or edit distance, to the above-mentioned combinatorial objects.

## 1. INTRODUCTION

We develop a version of Szemerédi's regularity lemma that is suitable for analyzing multicolorings of complete graphs and directed graphs. In proving our theorems we use as our guide the proof given by Alon, Fischer, Krivelevich and M. Szegedy [1] which proves a similar theorem in the case of graphs. Their idea is, when given a graph,  $G$ , they find an induced subgraph  $G'$  and two equipartitions,  $\mathcal{A}$  of  $V(G)$  and  $\mathcal{A}'$  of  $V(G')$ . The partitions  $\mathcal{A}$  and  $\mathcal{A}'$  have the same number of parts. Each part of  $\mathcal{A}'$  is large and contained in some part of  $\mathcal{A}$ , each pairwise density of the parts in  $\mathcal{A}'$  is close to the density of the corresponding pair in  $\mathcal{A}$ , but **all** pairs in  $\mathcal{A}'$  are regular. Our goal is to find an induced copy of  $H$  in  $G$ . If enough of the pairs of parts in  $\mathcal{A}$  have a sufficiently large density, we can apply the regularity lemma and Ramsey's theorem inside each of the parts of  $\mathcal{A}'$ . A slicing lemma ensures that the resulting subclusters (we call them miniclusters) are ready to witness the embedding of a graph  $H$ .

In fact, this approach works for any combinatorial object that has a sufficiently similar type of regularity lemma.

**Outline of the paper:** In the following subsections, we give basic definitions and the results we need for the graph version (Section 1.1), the multicolor version (Section 1.2) and the digraph version (Section 1.3). Section 1.4 gives the main result. In Section 2, we prove our main results for multicolored graphs and for directed graphs simultaneously – the main machinery depends very little on the combinatorial object to be studied. In Section 3, we apply our result to a specific problem related to edit distance.

**Definition 1.1.** A partition  $\mathcal{A} = \{V_i : 1 \leq i \leq k\}$  is an **equipartition** of a finite set if  $|V_i|$  and  $|V_{i'}|$  differ by at most 1 for all  $1 \leq i < i' \leq k$ . A **refinement** of  $\mathcal{A}$  is a partition  $\mathcal{B} = \{V_{i,j_i} : 1 \leq i \leq k, 1 \leq j_i \leq \ell_i\}$  such that  $V_i = \bigcup_{j=1}^{\ell_i} A_{i,j_i}$  for  $i = 1, \dots, k$ . The number of parts of a partition is its **order**.

Just to ensure some technicalities, we prove that every equipartition can be refined into an equipartition.

**Proposition 1.2.** Let  $\mathcal{A} = \{V_i : 1 \leq i \leq k\}$  be an equipartition of a finite set and let  $\ell$  be a positive integer,  $\ell \leq |V_i|$ ,  $i = 1, \dots, k$ . There exists a refinement of  $\mathcal{A}$  into  $k\ell$  parts that is an equipartition.

*Proof.* If all the  $V_i$  are the same size, it is clear that equipartitioning each will result in the equipartition we seek. Suppose the sizes of each  $V_i$  are  $s$  and  $s - 1$  such that  $s = q\ell + r$  for  $r \in \{0, \dots, \ell - 1\}$ . It suffices to show  $\lceil s/\ell \rceil$  and  $\lfloor (s - 1)/\ell \rfloor$  differ by at most one.

If  $r \neq 0$ , then  $\lceil s/\ell \rceil = q + 1$  and  $\lfloor (s - 1)/\ell \rfloor = q$ . If  $r = 0$ , then  $\lceil s/\ell \rceil = q$  and  $\lfloor (s - 1)/\ell \rfloor = q - 1$ .  $\square$

---

2010 *Mathematics Subject Classification.* Primary 05C35; Secondary 05C80.

*Key words and phrases.* edit distance, hereditary properties, localization, split graphs, colored regularity graphs.

This author's research partially supported by NSF grant DMS-0901008 and NSA grant H-98230-09-1-0063.

This author's research partially supported by NSF grant DMS-0901008 and by an Iowa State University Faculty Professional Development grant.

**1.1. Graph version.** A graph  $G$  is a pair  $(V, E)$  where  $V$  is a finite vertex set and  $E \subseteq \binom{V}{2}$ .

For disjoint vertex sets  $V_i, V_j$ , we denote  $e(V_i, V_j)$  to be number of edges with one endpoint in  $V_i$  and the other in  $V_j$ . The **density** of  $(V_i, V_j)$  is

$$d(V_i, V_j) := \frac{e(V_i, V_j)}{|V_i||V_j|}.$$

The **density vector** of the pair  $(V_i, V_j)$  is simply

$$\mathbf{d}(V_i, V_j) := (d(V_i, V_j), 1 - d(V_i, V_j)).$$

We say the pair  $(V_i, V_j)$  is a  $\gamma$ -**regular pair** if  $V'_i \subset V_i$  and  $V'_j \subset V_j$  such that both  $|V'_i| \geq \gamma|V_i|$  and  $|V'_j| \geq \gamma|V_j|$ , then  $|d(V'_i, V'_j) - d(V_i, V_j)| \leq \gamma$ .

A partition  $(V_1, \dots, V_k)$  of the vertex set of  $G$ , a graph on  $n$  vertices, is said to be a  $\gamma$ -**regular partition** if each of the following holds:

- $||V_i| - |V_j|| \leq 1$  for all  $i, j \in \{1, \dots, k\}$ .
- All but at most  $\gamma k^2$  of the pairs  $(V_i, V_j)$ ,  $1 \leq i < j \leq k$  are  $\gamma$ -regular.

A version of Szemerédi's lemma says the following:

**Theorem 1.3** (Szemerédi [6]). *For every  $m$  and  $\epsilon > 0$ , there exists an integer  $M = M(m, \epsilon)$  with the following property.*

*If  $G$  is a graph with  $n \geq M$  vertices, and  $\mathcal{A}$  is an equipartition of the vertex set of  $G$  of order at most  $m$ , then there exists a refinement  $\mathcal{B}$  of  $\mathcal{A}$  of order  $k$ , where  $m \leq k \leq M$ , which is  $\epsilon$ -regular.*

There are two important lemmas cited by Alon, et al. [1] which permit discussion of graph embedding. They have been presented and reproven many times, we give the statements here. The titles “Slicing lemma” and “Embedding lemma” can be found in the literature.

**Lemma 1.4** (Slicing lemma). *If  $(A, B)$  is a  $\gamma$ -regular pair with density  $\delta$  and  $A' \subset A$  and  $B' \subset B$  satisfy  $|A'| \geq \epsilon|A|$  and  $|B'| \geq \epsilon|B|$  for some  $\epsilon \geq \gamma$ , then  $(A', B')$  is a  $(\max\{2, \epsilon^{-1}\}\gamma)$ -regular pair with density at least  $\delta - \gamma$  and at most  $\delta + \gamma$ .*

**Lemma 1.5** (Embedding lemma). *For every  $0 < \eta < 1$  and positive integer  $k$  there exist  $\gamma = \gamma_{1.5}(\eta, k)$  and  $\delta = \delta_{1.5}(\eta, k)$  with the following property.*

*Suppose that  $H$  is a graph with vertices  $v_1, \dots, v_k$ , and that  $V_1, \dots, V_k$  is a  $k$ -tuple of disjoint vertex sets such that, for every  $1 \leq i < i' \leq k$ , the pair  $(V_i, V_{i'})$  is  $\gamma$ -regular, with density at least  $\eta$  if  $v_i v_{i'}$  is an edge of  $H$  and with density at most  $1 - \eta$  if  $v_i v_{i'}$  is not an edge of  $H$ . Then, at least  $\delta \prod_{i=1}^k |V_i|$  of the  $k$ -tuples  $w_1 \in V_1, \dots, w_k \in V_k$  span (induced) copies of  $H$  where each  $w_i$  plays the role of  $v_i$ .*

**1.2. Multicolor graph version.** We call an  $r$ -**graph** on  $n$  vertices a pair  $(V, c)$  where  $V$  is a set of size  $n$  and  $c: \binom{V}{2} \rightarrow \{1, \dots, r\}$  is a function known as the **coloring** of the edge set.

For disjoint vertex sets  $V_i, V_j$  and a color  $\rho \in \{1, \dots, r\}$ , we denote  $e_\rho(V_i, V_j)$  to be number of edges with one endpoint in  $V_i$  and the other in  $V_j$  and with color  $\rho$ . The  $\rho$ -**density** of  $(V_i, V_j)$  is

$$d_\rho(V_i, V_j) := \frac{e_\rho(V_i, V_j)}{|V_i||V_j|}.$$

The **density vector** of the pair  $(V_i, V_j)$  is simply

$$\mathbf{d}(V_i, V_j) := (d_1(V_i, V_j), \dots, d_r(V_i, V_j)).$$

We say the pair  $(V_i, V_j)$  is a  $\gamma$ -**regular pair** if  $V'_i \subset V_i$  and  $V'_j \subset V_j$  such that both  $|V'_i| \geq \gamma|V_i|$  and  $|V'_j| \geq \gamma|V_j|$ , then  $|d_\rho(V'_i, V'_j) - d_\rho(V_i, V_j)| \leq \gamma$  for each  $\rho \in \{1, \dots, r\}$ . Equivalently,  $\|\mathbf{d}(V'_i, V'_j) - \mathbf{d}(V_i, V_j)\|_\infty \leq \gamma$ .

A partition  $(V_1, \dots, V_k)$  of the vertex set of  $G$ , an  $r$ -colored graph on  $n$  vertices, is said to be a  $\gamma$ -**regular partition** if each of the following holds:

- $||V_i| - |V_j|| \leq 1$  for all  $i, j \in \{1, \dots, k\}$ .
- All but at most  $\gamma k^2$  of the pairs  $(V_i, V_j)$ ,  $1 \leq i < j \leq k$  are  $\gamma$ -regular.

The multicolor version of Szemerédi's lemma can be easily derived from a proof outline by Komlós and Simonovits [5]:

**Theorem 1.6** (Szemerédi [6]). *Fix an integer  $r \geq 2$ . For every  $\epsilon > 0$ , and positive integer  $m$ , there exists an integer  $CM = CM(m, \epsilon)$  with the following property.*

*If  $G$  is an  $r$ -graph with  $n \geq CM$  vertices, and  $\mathcal{A}$  is an equipartition of the vertex set of  $G$  with an order not exceeding  $m$ , then there exists a refinement  $\mathcal{B}$  of  $\mathcal{A}$  of order  $k$ , where  $m \leq k \leq CM$  which is  $\epsilon$ -regular.*

The classical formulation of Szemerédi's regularity lemma provides only the existence of the  $\epsilon$ -regular partition. However, its proof implies the more precise refinement result we state as Theorem 1.6. In addition, the classical formulation of the lemma allows for an exceptional set of size at most  $\epsilon n$ . We can, however, apply the original formulation to the graph  $G$  with a smaller parameter than  $\epsilon$  and evenly distribute the vertices in the exceptional set among the other clusters to get the result with the given value of  $\epsilon$ .

Multicolored graphs have their own Slicing and Embedding lemmas:

**Lemma 1.7** (Slicing lemma). *If  $(A, B)$  is a  $\gamma$ -regular pair in an  $r$ -graph such that  $(A, B)$  has density vector  $(d_1, \dots, d_r)$  and  $A' \subset A$  and  $B' \subset B$  satisfy  $|A'| \geq \epsilon|A|$  and  $|B'| \geq \epsilon|B|$  for some  $\epsilon \geq \gamma$ , then  $(A', B')$  is a  $(\max\{2, \epsilon^{-1}\}\gamma)$ -regular pair with density vector  $\mathbf{d} = (A', B')$  such that  $|d_\rho(A, B) - d_\rho(A', B')| \leq \gamma$  for each  $\rho \in \{1, \dots, r\}$  (equivalently,  $\|\mathbf{d}(A, B) - \mathbf{d}(A', B')\|_\infty \leq \gamma$ ).*

*Proof.* Let  $\eta = \max\{2, \epsilon^{-1}\}\gamma$ . We may assume  $\eta < 1$ , otherwise the lemma is trivially true as all pairs are  $\eta$ -regular whenever  $\eta \geq 1$ . In order to verify the regularity of  $(A', B')$ , choose  $A'' \subset A'$  and  $B'' \subset B'$  such that  $|A''| \geq \eta|A'|$  and  $|B''| \geq \eta|B'|$ . Consequently,

$$|A''| \geq \eta|A'| \geq \eta\epsilon|A| = \max\{2\epsilon, 1\}\gamma|A| \geq \gamma|A|$$

and similarly,  $|B''| \geq \gamma|B|$ . By the  $\gamma$ -regularity of  $(A, B)$ , we know that  $(A'', B'')$  has density vector  $\mathbf{d}(A'', B'')$  such that  $\|\mathbf{d}(A, B) - \mathbf{d}(A'', B'')\|_\infty \leq \gamma$ . Moreover, since  $|A'| \geq |A''| \geq \gamma|A|$  and  $|B'| \geq |B''| \geq \gamma|B|$ , then  $\|\mathbf{d}(A, B) - \mathbf{d}(A', B')\|_\infty \leq \gamma$ . By the triangle inequality,

$$\|\mathbf{d}(A', B') - \mathbf{d}(A'', B'')\|_\infty \leq \|\mathbf{d}(A, B) - \mathbf{d}(A', B')\|_\infty + \|\mathbf{d}(A, B) - \mathbf{d}(A'', B'')\|_\infty \leq 2\gamma \leq \eta.$$

The arbitrary choice of  $A''$  and  $B''$  means that  $(A', B')$  is  $\eta$ -regular.  $\square$

**Lemma 1.8** (Embedding lemma). *For every  $0 < \eta < 1$  and positive integer  $k$  there exist  $\gamma = \gamma_{1.8}(\eta, k)$  and  $\delta = \delta_{1.8}(\eta, k)$  with the following property.*

*Fix an integer  $r \geq 2$ . Suppose that  $H = (\{v_1, \dots, v_k\}, c)$  is an  $r$ -graph. Let  $G$  be an  $r$ -graph. Let  $V_1, \dots, V_k$  be a  $k$ -tuple of disjoint vertex sets of  $G$  such that for every  $1 \leq i < i' \leq k$  the pair  $(V_i, V_{i'})$  is  $\gamma$ -regular, such that the density  $d_\rho(V_i, V_{i'}) \geq \eta$  if  $v_i v_{i'}$  is an edge of  $H$  with color  $\rho$ , for each  $\rho$ ,  $1 \leq \rho \leq k$ . Then, at least  $\delta \prod_{i=1}^k |V_i|$  of the  $k$ -tuples  $(w_1, \dots, w_k)$  with  $w_1 \in V_1, \dots, w_k \in V_k$  span copies of  $H$  where each  $w_i$  plays the role of  $v_i$ .*

*Note that the case of  $r = 2$  is the case of induced graphs in which edges are color 1 and nonedges are color 2.*

*Proof.* We note that  $r$  plays no role at all in the definitions of  $\gamma$  and  $\delta$ . This is because  $\eta$  is the parameter that ensures the proper density for all colors. We will choose  $\gamma_{1.8}(\eta, k) = \min\{(\eta/2)^{k-1}, (1/6)^{k-1}\}$ .

We proceed via induction on  $k$  to determine the value of  $\delta_{1.8}(\eta, k)$ . The case of  $k = 1$  is trivial and  $\delta_{1.8}(\eta, 1) = 1$  for all  $\eta$ . Let  $k \geq 2$  and suppose there is such a function  $\delta_{1.8}(\eta, k-1)$ . Let

$$(1) \quad \gamma = \min\{(\eta/2)^{k-1}, (1/6)^{k-1}\}.$$

Consider  $V_k$ . Call a vertex  $w_k \in V_k$  bad if, for some  $i \in \{1, \dots, k-1\}$ ,  $w_k$  has less than  $(\eta - \gamma)|V_i|$  edges of color  $\rho = c(v_k, v_i)$  incident to it with the other endpoint in  $V_i$ .

Assume that more than  $\gamma|V_k|$  vertices in  $V_k$  are bad and let  $V'_k$  be the set of bad vertices. Then,  $d_\rho(V'_k, V_i) < \frac{(\eta - \gamma)|V'_k||V_i|}{|V_k||V_i|} = \eta - \gamma$ . On the other hand  $d_\rho(V_k, V_i) \geq \eta$ . So  $|d_\rho(V'_k, V_i) - d_\rho(V_k, V_i)| > \gamma$ , contradicting the fact that  $(V_k, V_i)$  is  $\gamma$ -regular.

Thus, the number of bad vertices is at most  $\gamma|V_k|$ . Therefore, there are at most  $(k-1)\gamma|V_k| < |V_k|$  vertices that are bad with respect to some  $V_i$ ,  $1 \leq i \leq k-1$ . Let  $w_k \in V_k$  be a vertex that is not bad with respect to each  $V_i$ . Let  $\overline{V}_i \subset V_i$  be a set of  $\lceil(\eta - \gamma)|V_i|\rceil$  vertices  $w_i$  such that  $w_i w_k$  has the correct color; i.e., the color of  $v_i v_k$ .

By the Slicing Lemma, each pair  $(\overline{V}_i, \overline{V}_{i'})$  for  $1 \leq i < i' \leq k-1$  is  $(\max\{2, (\eta - \gamma)^{-1}\}\gamma)$ -regular. The pairs also have that  $d_\rho(V_i, V_{i'}) \geq \eta - \gamma$  if  $v_i v_{i'}$  is an edge of  $H$  with color  $\rho$ .

In order to apply the inductive hypothesis, we must verify that

$$(2) \quad \max \{2, (\eta - \gamma)^{-1}\} \gamma \leq \gamma_{1.8}(\eta - \gamma, k - 1) = \min \left\{ \left( \frac{\eta - \gamma}{2} \right)^{k-2}, \left( \frac{1}{6} \right)^{k-2} \right\}.$$

If  $\eta - \gamma \geq 1/2$ , then (1) gives that  $\gamma = (1/6)^{k-1}$  and (2) reduces to  $2\gamma \leq (1/6)^{k-2}$ , which is true for all  $k$ .  
If  $\eta - \gamma < 1/2$  and  $\eta - \gamma \geq 1/3$ , then (1) gives that  $\gamma \leq (1/6)^{k-1}$  and (2) reduces to

$$\frac{\gamma}{\eta - \gamma} \leq \left( \frac{1}{6} \right)^{k-2}.$$

This is true because  $\gamma/(\eta - \gamma) \leq 3\gamma = 3(1/6)^{k-1} = \frac{1}{2}(1/6)^{k-2}$ .

If  $\eta - \gamma < 1/3$ , since  $\gamma \leq (\eta/2)^{k-1}$ , (2) reduces to

$$\begin{aligned} \frac{\gamma}{\eta - \gamma} &\leq \left( \frac{\eta - \gamma}{2} \right)^{k-2} \\ 2^{k-2}\gamma &\leq (\eta - \gamma)^{k-1}. \end{aligned}$$

To verify this, see that

$$2^{k-2}\gamma \leq 2^{k-2}(\eta/2)^{k-1} = \frac{1}{2}\eta^{k-1}$$

and that

$$(\eta - \gamma)^{k-1} \geq (\eta - (\eta/2)^{k-1})^{k-1} = \eta^{k-1} \left( 1 - \frac{\eta^{k-2}}{2^{k-1}} \right)^{k-1} \geq \eta^{k-1} (1 - 2^{1-k})^{k-1}.$$

Some calculus shows that  $(1 - 2^{-x})^x$  is increasing for  $x \geq 1$  and so we have

$$(\eta - \gamma)^{k-1} \geq \eta^{k-1} (1 - 2^{1-2})^{2-1} = \frac{1}{2}\eta^{k-1},$$

as needed.

Now that we have verified that we can use the inductive hypothesis, we do so and see that the number of copies of  $H - v_k$  in  $(\overline{V}_1, \dots, \overline{V}_{k-1})$  is at least  $\delta_{1.8}(\eta - \gamma, k - 1) \prod_{i=1}^{k-1} |\overline{V}_i|$ . So the total number of copies of  $H$  is at least

$$\begin{aligned} &\delta_{1.8}(\eta - \gamma, k - 1) \prod_{i=1}^{k-1} |\overline{V}_i| \cdot (1 - (k - 1)\gamma) |V_k| \\ &\geq \delta_{1.8}(\eta - \gamma, k - 1) (\eta - \gamma)^{k-1} (1 - (k - 1)\gamma) \prod_{i=1}^k |V_i|. \end{aligned}$$

With  $\gamma = \min \{(\eta/2)^{k-1}, (1/6)^{k-1}\}$ , set  $\delta_{1.8}(\eta, k) = \delta_{1.8}(\eta - \gamma, k - 1) (\eta - \gamma)^{k-1} (1 - (k - 1)\gamma)$ , the conditions of the Embedding Lemma are satisfied.  $\square$

**1.3. Directed graph version.** A **digraph** is defined to be a pair  $(V, E)$  where  $V$  is a labeled vertex set,  $E \subseteq (V)_2$  and  $(V)_2$  denotes the set  $V \times V - \{(v, v) : v \in V\}$ . It is convenient for us to view this as a coloring. That is, a digraph is a pair  $(V, c)$  where  $c : (V)_2 \rightarrow \{\circ, -, \leftarrow, \rightarrow\}$  is a function known as the **partial orientation** of the edge set. It has the property that, for distinct  $v, w$ ,

- $c(v, w) = c(w, v)$  if and only if  $c(v, w) \in \{\circ, -\}$  and
- $c(v, w) = \rightarrow$  if and only if  $c(w, v) = \leftarrow$ .

For convenience, we denote  $\overleftrightarrow{\mathcal{A}} := \{\circ, -, \leftarrow, \rightarrow\}$ . Here we interpret the color  $c(v, w) = \circ$  to mean that neither  $(v, w)$  nor  $(w, v)$  are in  $E$ , the color  $c(v, w) = -$  to mean that both  $(v, w)$  and  $(w, v)$  are in  $E$  and the color  $c(v, w) = \rightarrow$  to mean that  $(v, w) \in E$  and  $(w, v) \notin E$ .

In the directed case, we have the same notions of  $\gamma$ -regular pairs as in the multicolor case. The **density vector** of the pair  $(V_i, V_j)$  is somewhat similar as well:

$$\mathbf{d}(V_i, V_j) := (d_{\circ}(V_i, V_j), d_{-}(V_i, V_j), d_{\leftarrow}(V_i, V_j), d_{\rightarrow}(V_i, V_j)).$$

However, in the directed case, the order makes a difference. Although  $d_{\rho}(A, B) = d_{\rho}(B, A)$  for  $\rho \in \{\circ, -\}$ , it is also the case that  $d_{\rightarrow}(A, B) = d_{\leftarrow}(B, A)$ .

Alon and Shapira give the following version of Szemerédi's lemma:

**Theorem 1.9** (Alon-Shapira [2]). *For every  $\epsilon > 0$  and positive integer  $m$ , there exists an integer  $DM = DM(m, \epsilon)$  with the following property.*

*If  $G$  is a digraph  $n \geq DM$  vertices, and  $\mathcal{A}$  is an equipartition of the vertex set of  $G$  with an order not exceeding  $m$ , then there exists a refinement  $\mathcal{B}$  of  $\mathcal{A}$  of order  $k$ , where  $m \leq k \leq DM$  which is  $\epsilon$ -regular.*

Digraphs have their own Slicing and Embedding lemmas:

**Lemma 1.10** (Slicing lemma). *If  $(A, B)$  is a  $\gamma$ -regular pair in a digraph such that  $(A, B)$  has density vector  $(d_{\circlearrowleft}, d_{\circlearrowright}, d_{\leftarrow}, d_{\rightarrow})$  and  $A' \subset A$  and  $B' \subset B$  satisfy  $|A'| \geq \epsilon|A|$  and  $|B'| \geq \epsilon|B|$  for some  $\epsilon \geq \gamma$ , then  $(A', B')$  is a  $(\max\{2, \epsilon^{-1}\}\gamma)$ -regular pair with density vector  $\mathbf{d}' := (d'_{\circlearrowleft}, d'_{\circlearrowright}, d'_{\leftarrow}, d'_{\rightarrow})$  such that  $|d_{\rho} - d'_{\rho}| \leq \gamma$  for each  $\rho \in \{\circlearrowleft, \circlearrowright, \leftarrow, \rightarrow\}$  (equivalently  $\|\mathbf{d}(A, B) - \mathbf{d}(A', B')\|_{\infty} \leq \gamma$ ).*

The proof is identical to the multicolor case, Lemma 1.7.

**Lemma 1.11** (Embedding lemma). *For every  $0 < \eta < 1$  and positive integer  $k$  there exist  $\gamma = \gamma_{1.11}(\eta, k)$  and  $\delta = \delta_{1.11}(\eta, k)$  with the following property.*

*Suppose that  $H$  is a digraph with vertices  $v_1, \dots, v_k$ , and that  $V_1, \dots, V_k$  is a  $k$ -tuple of disjoint vertex sets of  $G$  such that for every  $1 \leq i < i' \leq k$  the pair  $(V_i, V_{i'})$  is  $\gamma$ -regular, such that the density  $d_{\rho}(V_i, V_{i'}) \geq \eta$  if  $(v_i, v_{i'})$  is an edge of  $H$  with color  $\rho$ . Then, at least  $\delta \prod_{i=1}^k |V_i|$  of the  $k$ -tuples  $(w_1, \dots, w_k)$  with  $w_1 \in V_1, \dots, w_k \in V_k$  span (induced) copies of  $H$  where each  $w_i$  plays the role of  $v_i$ .*

Again, the proof is identical to the multicolor case, Lemma 1.7.

**1.4. Main results.** The statement of the main result (Theorem 1.12) can be made in general with the definitions above. Recall that  $\mathbf{d}(V_i, V_{i'})$  denotes the density vector of the pair  $(V_i, V_{i'})$ .

**Theorem 1.12** (Alon, et al. [1]). *Fix  $r \geq 2$ . For every  $m$  and function  $\mathcal{E}$  with  $\mathcal{E} : \mathbb{N} \rightarrow (0, 1)$ , there exist  $S = S_{1.12}(r, m, \mathcal{E})$  and  $\delta = \delta_{1.12}(r, m, \mathcal{E})$  with the following property:*

*If  $G$  is a graph [ $r$ -graph, digraph] with  $n \geq S$  vertices then there exist an equipartition  $\mathcal{A} = \{V_i : 1 \leq i \leq k\}$  of  $G$  and an induced subgraph [induced  $r$ -subgraph, induced subdigraph]  $G'$  of  $G$ , with an equipartition  $\mathcal{A}' = \{V'_i : 1 \leq i \leq k\}$  of the vertices of  $G'$  that satisfy:*

- $S \geq k \geq m$ .
- $V'_i \subset V_i$  for all  $i \geq 1$ , and  $|V'_i| \geq \delta n$ .
- In the equipartition  $\mathcal{A}'$ , **all** pairs are  $\mathcal{E}(k)$ -regular.
- All but at most  $\mathcal{E}(0) \binom{k}{2}$  of the pairs  $1 \leq i < i' \leq k$  are such that  $\|\mathbf{d}(V_i, V_{i'}) - \mathbf{d}(V'_i, V'_{i'})\|_{\infty} < \mathcal{E}(0)$ .

Our contribution is to prove the case for multicolored graphs and digraphs. Although the proof is quite similar to that of N. Alon, E. Fischer, M. Krivelevich and M. Szegedy [1], there are subtleties that need to be addressed.

## 2. PROOF OF THE MAIN RESULTS

There is a plethora of lemmas that are required to prove our main result. Lemma 2.2 is a consequence of the defect form of the Cauchy-Schwarz Inequality which is stated without proof and can be found in [6]. Corollary 2.3 is a direct consequence of Lemma 2.2. Lemma 2.4 is a refinement lemma that allows the induction to take place and Lemma 2.5 is the main lemma, of which our main result, Theorem 1.12 is a direct consequence.

First, we need a definition which, in the context of multicolorings of the complete graph, comes from [5].

**Definition 2.1.** *Given an equipartition  $\mathcal{A} = \{V_i : 1 \leq i \leq k\}$  of the vertex set of a multicolored graph [digraph], we define the **index of  $\mathcal{A}$**  as follows:*

$$\text{ind}(\mathcal{A}) = \frac{1}{k^2} \sum_{\rho} \sum_{1 \leq i < i' \leq k} d_{\rho}^2(V_i, V_{i'}),$$

where in the case of multicolored graphs,  $\rho$  runs over all colors and in the case of digraphs, the colors  $\rho$  run over the set of four “colors” in the set  $\mathcal{A} = \{\circlearrowleft, \circlearrowright, \leftarrow, \rightarrow\}$ .

Note also that  $\text{ind}(\mathcal{A}) = \frac{1}{k^2} \sum_{1 \leq i < i' \leq k} \sum_{\rho} d_{\rho}^2(V_i, V_{i'}) \leq \frac{1}{k^2} \sum_{1 \leq i < i' \leq k} \left( \sum_{\rho} d_{\rho}(V_i, V_{i'}) \right)^2 \leq \frac{1}{2}$ .

**Lemma 2.2.** For all sequences of nonnegative numbers  $X_1, \dots, X_n$ , if for some  $m$ ,  $1 \leq m < n$

$$\sum_{k=1}^m X_k = \frac{m}{n} \sum_{k=1}^n X_k + \alpha,$$

then

$$\sum_{k=1}^n X_k^2 \geq \frac{1}{n} \left( \sum_{k=1}^n X_k \right)^2 + \frac{\alpha^2 n}{m(n-m)}.$$

(observe that  $\alpha$  need not be positive).

**Corollary 2.3.** Suppose that  $A$  and  $B$  are two disjoint sets of vertices of a multicolored graph [digraph]  $G$ , and  $\{A_j : 1 \leq j \leq \ell\}$  and  $\{B_j : 1 \leq j \leq \ell\}$  are their two respective partitions to sets of **equal** sizes, such that, for some color  $\rho$ , at least  $\epsilon \ell^2$  of the possible  $j, j'$  satisfy  $|d_\rho(A, B) - d_\rho(A_j, B_{j'})| \geq \frac{1}{2}\epsilon$ . Then,

$$\sum_{1 \leq j, j' \leq \ell} d_\rho^2(A_j, B_{j'}) > \ell^2 \left( d_\rho^2(A, B) + \frac{1}{8}\epsilon^3 \right).$$

**Proof of Corollary 2.3.** Under the above conditions, either at least  $\frac{1}{2}\epsilon \ell^2$  of the pairs  $j, j'$  are such that  $d_\rho(A_j, B_{j'}) - d_\rho(A, B) \geq \frac{1}{2}\epsilon$ , or at least  $\frac{1}{2}\epsilon \ell^2$  are such that  $d_\rho(A_j, B_{j'}) - d_\rho(A, B) \leq -\frac{1}{2}\epsilon$ . We use Lemma 2.2 with  $n = \ell^2$ ,  $m = \frac{1}{2}\epsilon \ell^2$ , and  $\alpha$  satisfying  $|\alpha| \geq \frac{1}{4}\epsilon^2 \ell^2$ . Furthermore, we use the fact that all  $|A_j| = |A|/\ell$  and all  $|B_{j'}| = |B|/\ell$  to obtain

$$\sum_{1 \leq j, j' \leq \ell} d_\rho(A_j, B_{j'}) = \ell^2 d(A, B).$$

Applying Lemma 2.2 to the sequence  $\{d_\rho(A_j, B_{j'})\}_{1 \leq j, j' \leq \ell}$ , we obtain

$$\sum_{1 \leq j, j' \leq \ell} d_\rho^2(A_j, B_{j'}) \geq \ell^2 d_\rho^2(A, B) + \frac{\frac{1}{16}\epsilon^4 \ell^6}{\frac{1}{2}\epsilon \ell^2 (\ell^2 - \frac{1}{2}\epsilon \ell^2)} > \ell^2 \left( d_\rho^2(A, B) + \frac{1}{8}\epsilon^3 \right)$$

as required.  $\square$

**Lemma 2.4.** Suppose that  $\mathcal{A} = \{V_i : 1 \leq i \leq k\}$  and its refinement  $\mathcal{B} = \{V_{i,j} : 1 \leq i \leq k, 1 \leq j \leq \ell\}$  be vertex partitions of a graph  $G$ , satisfying  $\text{ind}(\mathcal{B}) - \text{ind}(\mathcal{A}) \leq \frac{1}{64}r\epsilon^4$  for some  $\epsilon$ , and that the number of vertices of the graph is  $n > 512\epsilon^{-4}rkl$ . Then, for all possible  $i < i'$  but at most  $\epsilon \binom{k}{2}$  of them,  $|d_\rho(V_i, V_{i'}) - d_\rho(V_{i,j}, V_{i',j'})| < \epsilon$  holds simultaneously for all  $\rho$ , for all but a maximum of  $\epsilon \ell^2$  of the possible  $j, j'$ .

**Proof of Lemma 2.4.** Supposing the contrary and assuming  $\epsilon < 1$  and  $k > 1$ , we show that the index of  $\mathcal{B}$  is larger than that of  $\mathcal{A}$  by more than  $\frac{1}{64}r\epsilon^4$ . If not all of the sets of  $\mathcal{B}$  are of exactly the same size, let  $V'_{i,j}$  be  $V_{i,j}$  for sets of the smaller size and  $V'_{i,j}$  be  $V_{i,j}$  minus an arbitrarily chosen vertex for sets of the larger size. Defining also  $V'_i = \bigcup_{1 \leq j \leq \ell} V'_{i,j}$ , we define two new partitions  $\mathcal{B}' = \{V'_{i,j} : 1 \leq i \leq k, 1 \leq j \leq \ell\}$  and  $\mathcal{A}' = \{V'_i : 1 \leq i \leq k\}$  of a large induced submulticolored graph [subdigraph] of  $G$  (for each of these new partitions all its sets are of the same size). The assumption on  $n$  implies that  $|d_\rho(V_i, V_{i'}) - d_\rho(V'_i, V'_{i'})| < \frac{1}{256}\epsilon^4$  and  $|d_\rho(V_{i,j}, V_{i',j'}) - d_\rho(V'_{i,j}, V'_{i',j'})| < \frac{1}{256}\epsilon^4$  hold for all  $i, j, i', j', \rho$ . In particular,  $|\text{ind}(\mathcal{A}) - \text{ind}(\mathcal{A}')| < \frac{1}{128}\epsilon^4$  and  $|\text{ind}(\mathcal{B}) - \text{ind}(\mathcal{B}')| < \frac{1}{128}\epsilon^4$  hold, and for more than  $\epsilon \binom{k}{2}$  of the possible  $i < i'$ , the inequality  $|d_\rho(V'_i, V'_{i'}) - d_\rho(V'_{i,j}, V'_{i',j'})| > \epsilon - \frac{2}{256}\epsilon^4 > \frac{1}{2}\epsilon$  holds for at least  $\epsilon \ell^2$  of the possible  $j, j'$ . Using Corollary 2.3, we obtain

$$\begin{aligned} \text{ind}(\mathcal{B}') &\geq \frac{1}{k^2 \ell^2} \sum_{\rho} \sum_{\substack{1 \leq i < i' \leq k \\ 1 \leq j, j' \leq \ell}} d_\rho^2(V'_{i,j}, V'_{i',j'}) \\ &> \frac{1}{k^2 \ell^2} \sum_{\rho} \left( \ell^2 \sum_{1 \leq i < i' \leq k} d_\rho^2(V'_i, V'_{i'}) + \epsilon \binom{k}{2} \ell^2 \frac{1}{8}\epsilon^3 \right) \geq \text{ind}(\mathcal{A}') + \frac{1}{32}r\epsilon^4. \end{aligned}$$

This implies  $\text{ind}(\mathcal{B}) - \text{ind}(\mathcal{A}) \geq \text{ind}(\mathcal{B}') - \text{ind}(\mathcal{A}') - \frac{2}{128}r\epsilon^4 > \frac{1}{64}r\epsilon^4$ , completing the proof.  $\square$

The main lemma is Lemma 2.5.

**Lemma 2.5.** *Fix a positive integer  $r$ . For every integer  $m$  and function  $\mathcal{E}$  with  $\mathcal{E} : \mathbb{N} \rightarrow (0, 1)$ , there exists a number  $S = S_{2.5}(r, m, \mathcal{E})$  with the following property.*

*If  $G$  is an  $r$ -graph [digraph] with  $n \geq S$  vertices, then there exists an equipartition  $\mathcal{A} = \{V_i : 1 \leq i \leq k\}$  and a refinement  $\mathcal{B} = \{V_{i,j} : 1 \leq i \leq k, 1 \leq j \leq \ell\}$  of  $\mathcal{A}$  that satisfy:*

- $|\mathcal{A}| = k \geq m$  but  $|\mathcal{B}| = k\ell \leq S$ .
- For all  $1 \leq i < i' \leq k$  but at most  $\mathcal{E}(0)\binom{k}{2}$  of them, the pair  $(V_i, V_{i'})$  is  $\mathcal{E}(0)$ -regular.
- For all  $1 \leq i < i' \leq k$  and all  $1 \leq j, j' \leq \ell$  but at most  $\mathcal{E}(k)\ell^2$  of them, the pair  $(V_{i,j}, V_{i',j'})$  is  $\mathcal{E}(k)$ -regular.
- All  $1 \leq i < i' \leq k$  but at most  $\mathcal{E}(0)\binom{k}{2}$  of them are such that for all  $1 \leq j, j' \leq \ell$  but at most  $\mathcal{E}(0)\ell^2$  of them  $|d_\rho(V_i, V_{i'}) - d_\rho(V_{i,j}, V_{i',j'})| < \mathcal{E}(0)$  holds for each  $\rho \in \{1, \dots, r\}$ .

*Proof.* We may assume that  $m > 1$  and that  $\mathcal{E}(\kappa)$  is monotone nonincreasing. For convenience, let  $\epsilon = \mathcal{E}(0)$ .

If we are in the case of a multicolored graph, fix a positive integer  $r$ , and using the function  $CM$  from Theorem 1.6, let

$$T^{(1)} = CM(r, m, \epsilon)$$

and for  $i > 1$ , we define by induction

$$T^{(i)} = CM(r, T^{(i-1)}, 2\mathcal{E}(T^{(i-1)})(T^{(i-1)})^{-2}).$$

If we are in the case of a digraph, and using the function  $DM$  from Theorem 1.9, let

$$T^{(1)} = DM(m, \epsilon)$$

and for  $i > 1$ , we define by induction

$$T^{(i)} = DM(T^{(i-1)}, 2\mathcal{E}(T^{(i-1)})(T^{(i-1)})^{-2}).$$

In either case, we show that  $S = 512r\epsilon^{-4}T^{(64r\epsilon^{-4}+1)}$  satisfies the required property.

Given  $G$ , define  $\mathcal{A}_1$  to be an equipartition of order at least  $m$  but not greater than  $T^{(1)}$ , such that all pairs but at most  $\epsilon\binom{|\mathcal{A}_1|}{2}$  of them are  $\epsilon$ -regular. Define by induction for  $i > 1$  the equipartition  $\mathcal{A}_i$  to be a refinement of  $\mathcal{A}_{i-1}$ , of order not greater than  $T^{(i)}$  such that all of the pairs but at most

$$2\mathcal{E}(T^{(i-1)})(T^{(i-1)})^{-2}\binom{|\mathcal{A}_i|}{2} \leq 2\mathcal{E}(T^{(i-1)})(|\mathcal{A}_{i-1}|)^{-2}\binom{|\mathcal{A}_i|}{2}$$

are  $2\mathcal{E}(T^{(i-1)})(T^{(i-1)})^{-2} < \mathcal{E}(T^{(i-1)})$ -regular. The refinements are guaranteed by the original regularity lemma, either Theorem 1.6 (in the multicolor case) or Theorem 1.9 (in the digraph case).

Let us now choose the minimum  $i$  such that  $\text{ind}(\mathcal{A}_i) - \text{ind}(\mathcal{A}_{i-1}) \leq \frac{1}{64}r\epsilon^4$ . There certainly exists such an  $1 < i \leq 64r^{-1}\epsilon^{-4} + 1$  since the indices of each partition in the series are all between 0 and 1. We set  $\mathcal{A} = \mathcal{A}_{i-1}$  and  $\mathcal{B} = \mathcal{A}_i$ , and appropriately  $k = |\mathcal{A}_{i-1}| = |\mathcal{A}|$  and  $\ell = k^{-1}|\mathcal{A}_i| = |\mathcal{A}|^{-1}|\mathcal{B}|$ . We claim that  $\mathcal{A}$  and  $\mathcal{B}$  are the required partitions.

It is clear that  $\mathcal{B}$  is a refinement of  $\mathcal{A}$  and that they both satisfy the requirements with regards to their respective orders. It is also clear (by the assumption  $\mathcal{E}(\kappa) \leq \mathcal{E}(0) = \epsilon$ ) that  $\mathcal{A}$  satisfies the requirement regarding the regularity of its pairs. Since all but at most  $2\mathcal{E}(k)k^{-2}\binom{k\ell}{2} < \mathcal{E}(k)\ell^2$  of all the pairs of  $\mathcal{B}$  are  $\mathcal{E}(k)$ -regular, the condition regarding the regularity of pairs of  $\mathcal{B}$  in the formulation of the lemma follows. Finally, Lemma 2.4 shows that most densities of the pairs of  $\mathcal{B}$  differ from the corresponding densities of the pairs of  $\mathcal{A}$  by less than  $\epsilon$ , as in the formulation of the last condition of this lemma.  $\square$

**Proof of Theorem 1.12.** We may assume  $\mathcal{E}(\kappa) \leq \mathcal{E}(0)$ . Set  $\epsilon = \mathcal{E}(0)$ . Define  $\mathcal{E}'$  by setting  $\mathcal{E}'(\kappa) = \min \left\{ \mathcal{E}(\kappa), \frac{1}{4}\epsilon, \frac{1}{2}(\kappa+2)^{-1} \right\}$ , set  $S = S_{2.5}(r, m, \mathcal{E}')$  and  $\delta = \frac{1}{2}(S_{2.5}(m, \mathcal{E}'))^{-1}$ . Use Lemma 2.5 on  $G$ , finding the appropriate partitions  $\mathcal{A} = \{V_i : 1 \leq i \leq k\}$  and  $\mathcal{B} = \{V_{i,j} : 1 \leq i \leq k, 1 \leq j \leq \ell\}$ .

Now choose randomly, independently and uniformly  $j_i$  such that  $1 \leq j_i \leq \ell$  for each  $1 \leq i \leq k$ . With probability more than  $1/2$ , all the pairs  $(V_{i,j_i}, V_{i',j_{i'}})$  are  $\mathcal{E}'(k)$ -regular. In fact, the probability that there is some pair that is not  $\mathcal{E}(k)$ -regular is at most  $\mathcal{E}(k)'\binom{k}{2}$ .

Moreover, the expected number of pairs  $1 \leq i \leq i' \leq k$  for which  $|d_\rho(V_i, V_{i'}) - d_\rho(V_{i,j_i}, V_{i',j_{i'}})| \geq \epsilon$  for some  $\rho$  is no more than  $\frac{1}{4}\epsilon \binom{k}{2} + \frac{1}{4}\epsilon \binom{k}{2} = \frac{1}{2}\epsilon \binom{k}{2}$ , by the choice of  $\mathcal{E}'$ , so with probability at least  $1/2$ , no more than  $\epsilon \binom{k}{2}$  of the pairs satisfy this.

Therefore, there exists a choice of  $j_1, \dots, j_k$  such that all pairs  $(V_{i,j_i}, V_{i',j_{i'}})$  are  $\mathcal{E}(k)$ -regular, and all but at most  $\epsilon \binom{k}{2}$  of them satisfy  $|d_\rho(V_i, V_{i'}) - d_\rho(V_{i,j_i}, V_{i',j_{i'}})| < \epsilon$  for all  $\rho \in \{1, \dots, r\}$ . Defining  $G'$  as the induced subgraph spanned by  $\bigcup_{1 \leq i \leq k} V_{i,j_i}$ , and  $\mathcal{A}'$  by setting  $V_i = V_{i,j_i}$  achieves the required result.  $\square$

### 3. APPLICATION

An important feature of editing is the notion of the palette. Colloquially, the **palette** is the set of colors to which an edge can be changed. For an  $r$ -graph, the palette is always the set  $\{1, \dots, r\}$ . Note that if  $r = 2$ , this is the case of simple graphs. So, we will not define the palette for  $r$ -graphs, only focusing on it for digraphs.

**Definition 3.1.** In the case of digraphs, we say that  $\mathcal{P} \subseteq \overleftrightarrow{\mathcal{A}}$  is a **palette** if either none or both of “ $\rightarrow$ ” and “ $\leftarrow$ ” are in  $\mathcal{P}$  and every digraph is a pair  $(V, c)$  where  $V$  is a vertex set and  $c : (V)_2 \rightarrow \mathcal{P}$  is a coloring of the edge set of a complete graph on  $|V|$  vertices. There are 5 possible nontrivial palettes:

- (0)  $\mathcal{P}_0 = \overleftrightarrow{\mathcal{A}}$  is the most general case.
- (1)  $\mathcal{P}_1 = \{-, \leftarrow, \rightarrow\}$  is the case of simple digraphs such that every pair of vertices has at least one arc between them.
- (2)  $\mathcal{P}_2 = \{\circ, \leftarrow, \rightarrow\}$  is the case of **oriented graphs**; that is, no pair of vertices has two arcs between them.
- (3)  $\mathcal{P}_3 = \{\circ, -\}$  is the case of simple, undirected graphs.
- (4)  $\mathcal{P}_4 = \{\leftarrow, \rightarrow\}$  is the case of **tournaments**.

Recall that the vector is of the form  $(p, q)$  where  $p, q \geq 0$  and  $0 \leq 1 - p - 2q$ . In the cases in which the palette is not  $\overleftrightarrow{\mathcal{A}}$ , the relevant density vector must be further restricted.

- (1) In the case of  $\mathcal{P}_1 = \{-, \leftarrow, \rightarrow\}$ , then  $p + 2q = 1$ .
- (2) In the case of  $\mathcal{P}_2 = \{\circ, \leftarrow, \rightarrow\}$ , then  $p = 0$  and  $q \leq 1/2$ .
- (3) In the case of  $\mathcal{P}_3 = \{\circ, -\}$ , then  $q = 0$  and  $p \leq 1$ . This is the  $r$ -graph case where  $r = 2$  or simply the case of undirected graphs. See [3] and [4].
- (4) In the case of  $\mathcal{P}_4 = \{\leftarrow, \rightarrow\}$ , then  $p = 0$  and  $1 - p - 2q = 0$ , so  $q = 1/2$ .

Our application is one of edit distance and it shows that  $r$ -types [dir-types] are used to lower bound the edit distance function. It turns out that, trivially, they upper bound the edit distance function.

**Definition 3.2.** An  $r$ -**type**,  $K$ , is a pair  $(U, \phi)$ , where  $U$  is a finite set of vertices and  $\phi : U \times U \rightarrow 2^{\{1, \dots, r\}} \setminus \emptyset$ , such that  $\phi(x, y) = \phi(y, x)$  and  $\phi(x, x) \neq \{1, \dots, r\}$ , for all  $x, y \in U$ . Informally, we will view an  $r$ -type as a complete graph with a coloring of both vertices and edges using subsets of  $\{1, \dots, r\}$ . The **sub- $r$ -type** of  $K$  induced by  $W \subseteq U$  is the  $r$ -type achieved by deleting the vertices  $U - W$  from  $K$ .

We say that an  $r$ -graph  $H = (V, c)$  of a complete graph **embeds in type**  $K = (U, \phi)$ , and write  $H \mapsto K$ , if there is a map  $\gamma : V \rightarrow U$  such that  $c(\{v, v'\}) = c_0$  implies  $c_0 \in \phi(\gamma(v), \gamma(v'))$ .

Types are defined in a slightly different way for digraphs.

**Definition 3.3.** Let  $\mathcal{P} \subseteq \overleftrightarrow{\mathcal{A}}$  be a palette. A  $\mathcal{P}$ -**dir-type** or simply **dir-type** where the palette is understood,  $K$ , is a pair  $(U, \phi)$ , where  $U$  is a finite set of vertices and  $\phi : U \times U \rightarrow 2^{\mathcal{P}} \setminus \emptyset$ , such that

- (1) for distinct  $x, y$  and  $\rho \in \{\circ, -\}$ ,  $\phi(x, y) \ni \rho$  if and only if  $\phi(y, x) \ni \rho$  and
- (2)  $\phi(x, y) \ni \rightarrow$  if and only if  $\phi(y, x) \ni \leftarrow$ .

Moreover, for all  $x \in U$ ,  $\phi(x, x)$  is a nonempty proper subset of  $\mathcal{P}$ . The **sub-dir-type** of  $K$  induced by  $W \subseteq U$  is the dir-type achieved by deleting the vertices  $U - W$  from  $K$ .

We say that a directed graph  $H = (V, c)$  **embeds in type**  $K = (U, \phi)$ , and write  $H \mapsto K$ , if there is a map  $\gamma : V \rightarrow U$  such that, for distinct  $u, u' \in U$ ,  $c(v, v') \in \phi(u, u')$  whenever  $\gamma(v) = u$  and  $\gamma(v') = u'$  and for  $u \in U$ , the following occurs: (1) if exactly one of  $\{\leftarrow, \rightarrow\}$  is in  $\phi(u, u)$ , then the oriented edges of



$\gamma^{-1}(u)$  are a subdigraph of a transitive tournament (2) if neither  $\leftarrow$  nor  $\rightarrow$  is in  $\phi(u, u)$ , then  $\gamma^{-1}(u)$  has no oriented edges, (3) if  $\bigcirc \notin \phi(u, u)$ , then  $\gamma^{-1}(u)$  has no nonedges and (4) if  $- \notin \phi(u, u)$ , then  $\gamma^{-1}(u)$  has no undirected edges.

We define the set of types  $\mathcal{K}(\mathcal{H})$  that we need to consider for this problem.

**Definition 3.4.** Let  $\mathcal{H}$  be a hereditary property of  $r$ -graphs [digraphs]. We use the notation  $\mathcal{F}(\mathcal{H})$  to be the minimal set of  $r$ -graphs [digraphs] such that  $\mathcal{H} = \bigcap_{H \in \mathcal{F}(\mathcal{H})} \text{Forb}(H)$ , where  $\text{Forb}(H)$  denotes the property of having no induced copy of  $H$ .

We also denote  $\mathcal{K}(\mathcal{H})$  to be the set of all  $r$ -types [dir-types],  $K$ , such that  $H \not\vdash K$  for all  $H \in \mathcal{F}(\mathcal{H})$ .

The  $f_K$  function is what we use to compute the edit distance.

**Definition 3.5.** For an  $r$ -type,  $K = (U, c)$  on  $k$  vertices, and a density vector  $\mathbf{p} = (p_1, \dots, p_r)$ , we define the function  $f_K(\mathbf{p})$  as follows: For  $\rho = 1, \dots, r$ , let the matrix  $\mathbf{A}_\rho$  be such that the  $(i, j)^{\text{th}}$  entry is 1 if  $c(u_i, u_j) \ni \rho$  and zero otherwise. If  $\mathbf{J}$  denotes the  $k \times k$  all-ones matrix,  $\mathbf{1}$  denotes the  $k \times 1$  all-ones vector, then

$$f_K(\mathbf{p}) = \frac{1}{k^2} \mathbf{1}^T \left( \mathbf{J} - \sum_{\rho=1}^r p_\rho \mathbf{A}_\rho \right) \mathbf{1}.$$

The  $f_K$  function is defined in a slightly different way for digraphs.

**Definition 3.6.** For a dir-type,  $K = (U, c)$  on  $k$  vertices, and a density vector  $\mathbf{p} = (p, q)$ , we define the function  $f_K(\mathbf{p})$  as follows: For  $\rho = \bigcirc, -$ , let the matrix  $\mathbf{A}_\rho$  be such that the  $(i, j)^{\text{th}}$  entry is 1 if  $c(u_i, u_j) \ni \rho$  and zero otherwise. The matrix  $\mathbf{A}_{\rightarrow}$  has the property that the  $(i, j)^{\text{th}}$  entry is

$$\begin{cases} 1, & \text{if } c(u_i, u_j) \text{ contains exactly one member of } \{\leftarrow, \rightarrow\}; \\ 2, & \text{if } c(u_i, u_j) \supseteq \{\leftarrow, \rightarrow\}; \text{ and} \\ 0, & \text{otherwise.} \end{cases}$$

If  $\mathbf{J}$  denotes the  $k \times k$  all-ones matrix,  $\mathbf{1}$  denotes the  $k \times 1$  all-ones vector, then

$$f_K(\mathbf{p}) = \frac{1}{k^2} \mathbf{1}^T (\mathbf{J} - (1 - p - 2q) \mathbf{A}_{\bigcirc} - p \mathbf{A}_{-} - q \mathbf{A}_{\rightarrow}) \mathbf{1}.$$

The entry of 2 is necessary in order to account for the fact that fewer editing operations are required if both directions are permitted rather than simply one direction.

Finally, some definitions with respect to edit distance:

**Definition 3.7.** For  $r$ -graphs [digraphs]  $G = (V, c)$  and  $G' = (V, c')$  on the same labeled vertex set, the expression  $\text{dist}(G, G')$  counts the number of pairs of vertices  $v, v'$  such that  $c(v, v') \neq c'(v, v')$ .

The distance of  $G$  from  $\mathcal{H}$  is  $\min\{\text{dist}(G, G') : G' \in \mathcal{H}\}$ .

We need to express the main application differently in the case of  $r$ -graphs and digraphs. However, only the  $r$ -graph version will be proven.

**Theorem 3.8.** Let  $G'$  be an  $r$ -graph in hereditary property  $\mathcal{H} = \bigcap_{H \in \mathcal{F}(\mathcal{H})} \text{Forb}(H)$  and  $\mathbf{p} = (p_1, \dots, p_r)$  be a probability vector. Then, there exists an  $r$ -type  $K \in \mathcal{K}(\mathcal{H})$  such that  $H \not\vdash K$  for all  $H \in \mathcal{F}(\mathcal{H})$  and with probability going to 1 as  $n \rightarrow \infty$ ,  $\text{dist}(G_{n, \mathbf{p}}, \mathcal{H}) \geq f_K(\mathbf{p}) \binom{n}{2} - o(n^2)$ .

**Theorem 3.9.** Let  $G'$  be a digraph in hereditary property  $\mathcal{H} = \bigcap_{H \in \mathcal{F}(\mathcal{H})} \text{Forb}(H)$  and  $\mathbf{p} = (p, q)$  be a probability vector. Then, there exists a dir-type  $K \in \mathcal{K}(\mathcal{H})$  such that  $H \not\vdash K$  for all  $H \in \mathcal{F}(\mathcal{H})$  and with probability going to 1 as  $n \rightarrow \infty$ ,  $\text{dist}(G_{n, \mathbf{p}}, \mathcal{H}) \geq f_K(\mathbf{p}) \binom{n}{2} - o(n^2)$ .

*Proof.* Fix  $\eta \gg \delta \gg \epsilon > 0$ . Let  $G$  be distributed according to  $G_{n, \mathbf{p}}$  and  $G' \in \mathcal{H}$  be a graph of distance  $\text{dist}(G, \mathcal{H})$  from  $G$ . Apply Theorem 1.12 with  $m = \epsilon^{-1}$  and any decreasing function  $\mathcal{E}$  for which  $\mathcal{E}(0) = \epsilon$  to  $G'$  and consider the partition  $\mathcal{A}' = (V'_1, \dots, V'_k)$ . Construct the  $r$ -type [dir-type]  $K_0 = (U, c_0)$  on vertex set  $U = \{u_1, \dots, u_k\}$  as follows. For distinct  $i, j$ ,  $c_0(u_i, u_j) \ni \rho$  if and only if the pair  $(V'_i, V'_j)$  is  $\mathcal{E}(k)$ -regular such that the color  $\rho$  occurs with density at least  $\delta$ .

Now, we shall define  $c_0$  on the vertices; i.e.,  $c_0(u_i, u_i)$ ,  $u_i \in U$ , such that  $K_0 \in \mathcal{K}(\mathcal{H})$ . Assume no such assignment to the vertices exists; i.e., for any choice of colors of the vertices, there exists an  $H \in \mathcal{F}(\mathcal{H})$

for which  $H \mapsto K_0$ . Apply the regularity lemma (Theorem 1.6 in the  $r$ -graph case or Theorem 1.9 in the digraph case) to each of the clusters  $V_i'$  and use Ramsey theory find a clique of miniclusters that are regular with positive density in the same color. Assign that color to  $u_i$  to complete the definition of  $K_0$ . Using the relevant slicing and embedding lemmas, we see that if  $H \mapsto K_0$ , then there is an induced copy of  $H$  in  $G'$ , a contradiction. (See the authors and Kézdy [4] for details in the graph case.)

As to counting the number of changes, for all distinct  $i < i'$ , it is the case that  $d_{G',\rho}(V_i, V_{i'}) = 0$  for all  $\rho \notin c_0(v_i, v_{i'})$ . By Theorem 1.12, we can look at the equipartition  $\mathcal{A}$  and see that for all but  $\mathcal{E}(0)k^2$  such pairs,  $d_{G',\rho}(V_i, V_{i'}) \leq \mathcal{E}(0)$  for all  $\rho \notin c_0(u_i, u_{i'})$ . Now consider the equipartition  $\mathcal{A}$  as applied to  $G$ . We see that

$$\begin{aligned} \text{dist}(G, G') &\geq \sum_{1 \leq i < i' \leq k} \sum_{\rho \notin c_0(u_i, u_{i'})} (d_{G,\rho}(V_i, V_{i'}) - \mathcal{E}(0)) |V_i| |V_{i'}| - \mathcal{E}(0)k^2 \left\lceil \frac{n}{k} \right\rceil^2 \\ &\geq \sum_{1 \leq i < i' \leq k} \sum_{\rho \notin c_0(u_i, u_{i'})} d_{G,\rho}(V_i, V_{i'}) \left\lceil \frac{n}{k} \right\rceil^2 - \mathcal{E}(0)r \binom{k}{2} \left\lceil \frac{n}{k} \right\rceil^2 - \mathcal{E}(0)k^2 \left\lceil \frac{n}{k} \right\rceil^2. \end{aligned}$$

A routine Chernoff bound computation shows that, since  $k$  is bounded, if  $1 \leq i < i' \leq k$ , then the probability that  $d_{G,\rho}(V_i, V_{i'}) < p_\rho - \lfloor n/k \rfloor^{-1/3}$  is at most  $\exp\{-2\lfloor n/k \rfloor^{1/3}\}$ . Given an equipartition of  $V$  of order  $k$ , the probability that there exists some pair  $(V_i, V_{i'})$  and some  $\rho \in \{1, \dots, r\}$  such that  $d_{G,\rho}(V_i, V_{i'}) < p_\rho - \lfloor n/k \rfloor^{-1/3}$  is at most  $r \binom{k}{2} \exp\{-2\lfloor n/k \rfloor^{1/3}\}$ . The number of equipartitions, disregarding the labeling of the vertices, is bounded by a function of  $S = S_{2.5}(r, \mathcal{E}(0)^{-1}, \mathcal{E})$ . Hence, the probability of having an equipartition with one such pair is  $O(\exp\{-2(n/S)^{1/3}\})$ .

So, with that probability, and the fact that  $\mathcal{E}(0) = \epsilon$ ,

$$\begin{aligned} \text{dist}(G, G') &\geq \sum_{1 \leq i < i' \leq k} \sum_{\rho \notin c_0(u_i, u_{i'})} p_\rho \left\lceil \frac{n}{k} \right\rceil^2 - \left\lceil \frac{n}{k} \right\rceil^{5/3} r \binom{k}{2} - \epsilon r \binom{k}{2} \left\lceil \frac{n}{k} \right\rceil^2 - \epsilon k^2 \left\lceil \frac{n}{k} \right\rceil^2 \\ &= \frac{1}{2} \sum_{\substack{1 \leq i, i' \leq k \\ i \neq i'}} \sum_{\rho \notin c_0(u_i, u_{i'})} p_\rho \left\lceil \frac{n}{k} \right\rceil^2 - \left\lceil \frac{n}{k} \right\rceil^{5/3} r \binom{k}{2} - \epsilon r \binom{k}{2} \left\lceil \frac{n}{k} \right\rceil^2 - \epsilon k^2 \left\lceil \frac{n}{k} \right\rceil^2 \\ &\geq \frac{1}{2} \sum_{1 \leq i, i' \leq k} \sum_{\rho \notin c_0(u_i, u_{i'})} p_\rho \left\lceil \frac{n}{k} \right\rceil^2 - \frac{k}{2} \left\lceil \frac{n}{k} \right\rceil^2 - r \binom{k}{2} \left\lceil \frac{n}{k} \right\rceil^{5/3} - \epsilon r \binom{k}{2} \left\lceil \frac{n}{k} \right\rceil^2 - \epsilon k^2 \left\lceil \frac{n}{k} \right\rceil^2 \\ &= f_K(\mathbf{p}) \frac{k^2}{2} \left\lceil \frac{n}{k} \right\rceil^2 - \frac{k}{2} \left\lceil \frac{n}{k} \right\rceil^2 - r \binom{k}{2} \left\lceil \frac{n}{k} \right\rceil^{5/3} - \epsilon r \binom{k}{2} \left\lceil \frac{n}{k} \right\rceil^2 - \epsilon k^2 \left\lceil \frac{n}{k} \right\rceil^2 \\ &\geq f_K(\mathbf{p}) \binom{n}{2} \\ (3) \quad &- \left[ \left( \binom{n}{2} - \frac{k^2}{2} \left\lceil \frac{n}{k} \right\rceil^2 \right) + \frac{k}{2} \left\lceil \frac{n}{k} \right\rceil^2 + r \binom{k}{2} \left\lceil \frac{n}{k} \right\rceil^{5/3} + \epsilon r \binom{k}{2} \left\lceil \frac{n}{k} \right\rceil^2 + \epsilon k^2 \left\lceil \frac{n}{k} \right\rceil^2 \right]. \end{aligned}$$

Since  $k \geq m = \epsilon^{-1}$ , we can see that the error term in (3) is  $O(\epsilon n^2)$ . So, for any  $\eta > 0$ , the probability that  $\text{dist}(G_{n,\mathbf{p}}, \mathcal{H}) \geq f_K(\mathbf{p}) \binom{n}{2} - \eta n^2$  goes to 1 as  $n \rightarrow \infty$ .  $\square$

## REFERENCES

- [1] N. Alon, E. Fischer, M. Krivelevich and M. Szegedy, Efficient testing of large graphs, *Combinatorica* **20**(4) (2000), no. 4, 451–476.
- [2] N. Alon and A. Shapira, Testing subgraphs in directed graphs, *J. Comput. System Sci.* **69** (2004), no. 3, 353–382.
- [3] N. Alon and U. Stav, What is the furthest graph from a hereditary property? *Random Structures Algorithms* **33** (2008), no. 1, pp. 87–104.
- [4] M. Axenovich, A. Kézdy and R. Martin, On the editing distance of graphs, *J. Graph Theory* **58**(2) (2008), pp. 123–138.
- [5] J. Komlós and M. Simonovits, Szemerédi’s regularity lemma and its applications in graph theory, *Combinatorics, Paul Erdős is eighty, Vol. 2 (Keszthely, 1993)*, 295–352, Bolyai Soc. Math. Stud., 2, János Bolyai Math. Soc., Budapest, 1996.
- [6] E. Szemerédi, Regular partitions of graphs, *Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976)*, pp. 399–401, Colloq. Internat. CNRS, 260, CNRS, Paris, 1978.

DEPARTMENT OF MATHEMATICS, IOWA STATE UNIVERSITY, AMES, IOWA 50011  
*E-mail address:* **axenovic@iastate.edu**

DEPARTMENT OF MATHEMATICS, IOWA STATE UNIVERSITY, AMES, IOWA 50011  
*E-mail address:* **rymartin@iastate.edu**